Usable speech detection using a context dependent Gaussian mixture model classifier

Robert E Yantorno, Temple University
Brett Y Smolenski
Ananth N Iyer
Jashmin K Shah
USABLE SPEECH DETECTION USING A CONTEXT DEPENDENT GAUSSIAN MIXTURE MODEL CLASSIFIER

Robert E. Yantorno, Brett Y. Smolenski, Ananth N. Iyer, Jashmin K. Shah
Temple University/ECE Dept. 12th & Norris Streets, Philadelphia, PA 19122-6077, USA
robert.yantorno@temple.edu, bsmolens@temple.edu, aniyer@temple.edu, shah@temple.edu
http://www.temple.edu/speech_lab

ABSTRACT

Speech that is corrupted by nonstationary interference, but contains segments that are still usable for applications such as speaker identification or speech recognition, is referred to as “usable” speech. A common example of nonstationary interference occurs when there is more than one person talking at the same time, which is known as co-channel speech. In general, the above speech processing applications do not work in co-channel environments; however, they can work on the extracted usable segments. Unfortunately, currently available usable speech measures only detect about 75% of the total available usable speech. The first reason for this high error stems from the fact that no single feature can accurately identify all the usable speech characteristics. This situation can be resolved by using a Gaussian Mixture Model (GMM) based classifier to combine several usable speech features. A second source of error stems from the fact that the current usable speech measures treat each frame of co-channel data independently of the decisions made on adjacent frames. This situation can be resolved when a Hidden Markov Model (HMM) is used to incorporate any context dependent information in adjacent frames. Using this approach we were able to obtain 84% detection of usable speech with a 16% false alarm rate.

1. INTRODUCTION

In the field of audio restoration, such as the removing of clicks in gramophone recordings, it is common practice to first detect and remove the damaged portions of the signal and then interpolate the removed sections [1]. This approach has the advantage of only processing the damaged portions of the signal. It is interesting to note that this approach to audio restoration is commonly used when dealing with nonstationary interference, but has not yet been applied to the nonstationary interference in co-channel speech.

In addition, even for speech having little interference there exists several common situations like coughing, yawning, laughing, etc., that lie outside the training sets of typical speech processing applications. Further, most speech processing systems use feature vectors derived from physical models of the speech production process, such as the Linear Prediction Coefficients (LPC), which assume an all-pole model of the vocal tract [2]. However, for some classes of phonemes, such as nasals, it is known that the underlying production mechanisms are best described using pole-zero models [3]. It is unlikely that the above mentioned kinds of data would be very useful when input to a speech processing application, and hence, one would not want to processes these segments. The situation is similar to when the statistician identifies and removes outliers or when one chooses not to process low energy and unvoiced speech frames.

The traditional approach to processing highly corrupted speech has been to enhance the speech while attenuating the interference [4]. However, recently a novel approach to co-channel speech processing has been proposed [5]. Like the audio restoration approach to click removal, the portions of the co-channel speech that are highly corrupted are first detected and removed. Within a co-channel utterance, where both speakers are contributing the same overall energy, there exist several segments of speech where one of the speakers is 15 dB or more above the other speaker [5]. It has been shown that when the target speaker is at least 15 dB greater than the interfering speaker, 80% reliable identification of the target speaker can be obtained [6]. Hence, these segments with a high Target-to-Interferer Ratio (TIR) may be considered usable with respect to speaker identification. The TIR was computed by taking the value, in dB, of the ratio of signal power to interferer power. Since for speaker identification it is not necessary to make a decision on every frame of data, the system can be implemented in a co-channel environment by extracting out and processing only the usable segments. Fortunately, current research has shown that about 35% of a co-channel utterance is usable speech [6].

Recent advances in co-channel speech processing have produced several usable speech measures, which yield some indication of the TIR [7] [8] [9] [10]. Such measures are necessary to determine usability in an operational environment, since a priori knowledge of the TIR would not be available. Unfortunately, currently available usable speech measures only detect about 75% of the total available usable speech. One reason for this high detection error stems from the fact that the measures treat each frame of co-channel data independently of the decisions made on adjacent frames. Another source of error stems from the fact that no single usable speech measure is capable of identifying all of the characteristics of usable speech [11]. It is the goal of this research to increase the performance of usable speech identification by combining several measures and making the classification process context dependant.

The system in Figure 1 (below) illustrates the approach taken. The features used in this research were Linear Prediction Coefficients (LPC) along with a linear discriminant (LD) based feature derived from the LPC residual. The features were then orthogonalized, to make them independent, and passed through a GMM classifier. Previously proposed approaches used the much less sophisticated classification techniques of nonlinear estimation and Quadratic Discriminant Analysis (QDA) on only two usable speech measures, which did not make use of contextual information [11] [12]. The decisions of the GMM are
then passed through a Maximum Likelihood Sequence Detector to determine the most probable sequence of usable and unusable states given the output of the output of the GMM.

**Figure 1:** Block diagram of context dependent usable speech classifier.

Using a GMM type classifier is used, the desirable features are those that have a distribution well modeled by a mixture of Gaussians. Although the actual distribution of the LPC’s of a speaker for a particular phoneme may not be Gaussian distributed, the estimate of them is [13]. When one includes the estimates of the LPC’s across several phonemes, a mixture of Gaussians should result regardless of the orthogonalization stage. In addition, Linear Discriminant Analysis (LDA) is used on the LPC residual to yield an additional novel feature that incorporates any remaining useful information. Further research using additional nonlinear features having the above desirable properties is currently ongoing.

2. BACKGROUND

2.1 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) was used in an attempt to capture all the remaining information left in the LPC residual using one additional feature. The goal of linear discriminant analysis is to use a linear transformation to project the set of raw testing data vectors onto a vector space of lower dimension such that some metric of class discrimination is maximized [14]. The metric most often used is the ratio of the between class scatter (variance) to the within class scatter:

\[
trace\left(S_w^{-1}S_b\right)
\]  

(1)

The result of this minimization for the two class (usable or unusable) problem is the following linear transformation (matrix equation):

\[
y = (\mu_1 - \mu_2)^T S_w^{-1} x
\]  

(2)

where the \(\mu\) are the two mean vectors of the two class’s data vectors and \(x\) is the data test vector [14]. The mean vectors and within-class-scatter matrix were estimated using the sample mean and sample variance of the training vectors.

This transformation produces the 1-dimensional feature \(y\) from the LPC-residual data frames, which for this research were 80 samples (10ms frame at 8kHz sampling rate) in length. Hence, the transformation is from \(\mathbb{R}^{80}\) to \(\mathbb{R}\). Since the feature generated by this approach is a linear combination of a large number of independent identically distributed random variables, the feature’s probability distribution will be highly Gaussian regardless of the distribution of \(x\) [15]. Exploring other metrics as well as nonlinear transformations is currently ongoing.

2.2 Gaussian Mixture Model

Formally a random vector \(x\) that is described by a mixture of \(M\) Gaussians has a probability density function of the form:

\[
f(x) = \sum_{i=1}^{M} \lambda_i N(\mu_i, \Sigma_i)
\]  

(3)

where the \(N(\mu, \Sigma)\) are multivariate Gaussian distributions having mean vector \(\mu\) and covariance matrix \(\Sigma\) [16]. The \(\lambda\) sum to one and indicate the relative weight of each Gaussian component in the mixture. It can be shown that any distribution can be approximated with arbitrary precision using a mixture of enough Gaussians [14].

To obtain the parameters \(\lambda\), \(\mu\), and \(\Sigma\), the Expectation Maximization (EM) algorithm was used, which is an iterative implementation of maximum likelihood estimation using incomplete information about the underlying probability distributions [16]. Eight mixture components (\(M=64\)) were used, since this amount produced the lowest detection error.

In general, each covariance matrix in the mixture contains \(N^2\) elements (where \(N\) is the dimension of the feature vector) that need to be estimated. If the features are chosen such that they are independent, than all the off-diagonal elements of the covariance matrices will be zero [15]. Hence, one would like to use independent features, since only \(N\) parameters would need to be estimated.

2.3 Hidden Markov Model

In order to make the classifier context dependent, it would be helpful to use a statistical model that exploits as much a priori information about the TIR as possible [14]. One challenge regarding this is that the segmental TIR process is a nonstationary process. To model the non-stationary aspects of the TIR, the following HMM is proposed, Figure 2 (below).

**Figure 2:** State diagram of the HMM process of co-channel speech frames.

We say the model is ‘hidden’ because one cannot observe the actual states, just the statistical characteristics of the signal for a particular state [17]. For the usable state, one person is talking with little interference. For the unusable state, both talkers are contributing about equal energy. Hence, the transition probabilities of this process are related to the statistics of the silent portions of speech. Each state corresponds to a 40ms frame of the co-channel signal and, hence, the signal is quasi-stationary in this time interval [2].

The state transition matrix \(T\) of this process is:

\[
T = \begin{bmatrix}
p & 1-p \\
1-q & q
\end{bmatrix}
\]  

(4)

where \(p\) is the probability of the next frame being usable given that the current frame is usable, and \(q\) is the probability of the
next frame being unusable given that the current frame is unusable. These probabilities were estimated using the measure’s training data. One can notice that this model will only make use of dependence between adjacent frames. Fortunately, current research has shown that little dependence exists between anything but adjacent frames [5]. Using the state transition matrix in conjunction with the celebrated Forward-Backward algorithm it is possible to determine the maximum likelihood sequence of states given the output of the GMM classifier [17].

3. METHODS
For this research 10 single speaker utterances, 5 male and 5 female, were taken from the TIMIT speech corpus. These 10 utterances were used to form a co-channel speech database of 45 co-channel files (10 choose 2 = 45). The files were down sampled to 8kHz and the longer file in each pair was truncated to make both files the same length. The files were then combined at 0 dB overall TIR to form the co-channel utterance. To control the variability and eliminate any bias between the dialect regions, only one dialect region was used (region 1 of the TIMIT data base).

It should be noted that in an operational environment it is highly unlikely that two speakers would be talking over each other during the entire utterance. In addition, each utterance would not have exactly the same length or have the same energy. The reason for using this approach was to capture the worst possible scenario, with respect to both speakers, that one could expect in a co-channel environment.

Once the co-channel utterance was formed it was broken down into 40 ms frames with no overlap, since it has been demonstrated that speaker identification reliability has little dependence on overlap [6]. For each frame, the values for the features, TIR, signal energy, and spectral flatness were computed. Signal energy and spectral flatness were necessary in order to exclude silence and unvoiced frames, since usable speech measures would not be used with these frames. Usable speech measures are designed to work with only voiced speech, since unvoiced frames provide little information useful for speaker identification [6].

Training data was used for obtaining the parameters of the GMM classifier and MLSE detector. Once these models were obtained, it was possible to use testing data to classify what frames of the co-channel speech were usable ([TIR] > 15dB) and those frames that were unusable ([TIR] < 15dB). The absolute value is necessary, since usable speech can come from both speakers. Half of the 45 co-channel speech files (22) were randomly selected to train the system. The remaining half (23 co-channel files) were used for testing.

4. RESULTS
Figure 3 (below) shows the classification results for the APPC measure alone, quadratic discriminant analysis (QDA) classifier using APPC and SAPVR-residual measures as features, context independent GMM classifier using the LPC-based features, and the context dependent GMM classifier. Since the minimum probability of error criterion is used in determining the decision boundary surface of the classifier, the percent of Misses (%Misses = 100% - %Hits) equals the percentage of false alarms. However, one can easily choose to weight the false alarms differently than the misses and obtain a different decision boundary surface. The context dependent GMM classifier was able to obtain 84% detection of usable speech with a 16% false alarm rate. This amounts to a 38% reduction in total detection error over the APPC measure alone.

Figure 3: Percent hits and False Alarms (FA) for APPC measure, QDA classification using APPC and SAPVR-residual measures, context independent GMM classifier using LPC-based features, and context dependent (CD) GMM classifier.

5. FURTHER RESEARCH
More usable features need to be developed. Some current candidates include using pole-zero parameters, sinusoidal model parameters, as well as, nonlinear features of the vocal tract such as those derived from the Teager energy operator [18]. Further, parameters derived from the glottis such as the Liljencrants-Fant model of the glottal flow derivative, may give an indication of the quality of the speech [19]. In addition, the use of other classification techniques such as Support Vector Machines (SVM) has yet to be explored.

The current approach to usable speech segmentation is to partition the signal into short fixed length frames with no overlap. Segmentation is always necessary, since the speech signal is nonstationary [2]. However, a more intelligent approach to segmenting the speech signal would be to identify the stationary regions in the speech signal and process those entire segments. Iterative feature extraction and sequential usable speech detection should improve the resolution capabilities of the classifier as well as make it context dependent by default.

Also, usable speech can be defined with respect to the intended application, as opposed to the TIR value, by studying what types of frames work with the system. In addition to improving speaker identification systems, several other applications of usable speech are currently under development including a speaker count and speaker separation system [20].

ACKNOWLEDGEMENT
This effort was sponsored by the Air Force Research Laboratory, Air Force Material Command, and USAF, under agreement number F30602-02-2-0501. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon.

DISCLAIMER
The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily the official policies or endorsements, either expressed or implied, of the Air Force Research Laboratory, or the U.S. Government.
6. REFERENCES


